Single-Trial Classification of Event-Related Potentials in Rapid Serial Visual Presentation Tasks Using Supervised Spatial Filtering

Hubert Cecotti, Miguel P. Eckstein, and Barry Giesbrecht

***Abstract*— Accurate detection of single-trial event-related potentials (ERPs) in the electroencephalogram (EEG) is a difficult problem that requires efficient signal processing and machine learning techniques. Supervised spatial filtering methods that enhance the discriminative information in EEG data are commonly used to improve single-trial ERP detection. We propose a convolutional neural network (CNN) with a layer dedicated to spatial filtering for the detection of ERPs and with training based on the maximization of the area under the receiver operating characteristic curve (AUC). The CNN is compared with three common classifiers: 1) Bayesian linear discriminant analysis; 2) multilayer perceptron (MLP); and 3) support vector machines. Prior to classification, the data were spatially filtered with xDAWN (for the maximization of the signal-to-signal-plus-noise ratio), common spatial pattern, or not spatially filtered. The 12 analytical techniques were tested on EEG data recorded in three rapid serial visual presentation experiments that required the observer to discriminate rare target stimuli from frequent nontarget stimuli. Classification performance discriminating targets from nontargets depended on both the spatial filtering method and the classifier. In addition, the nonlinear classifier MLP outperformed the linear methods. Finally, training based AUC maximization provided better performance than training based on the minimization of the mean square error. The results support the conclusion that the choice of the systems architecture is critical and both spatial filtering and classification must be considered together.**

***Index Terms*— Brain–computer interface (BCI), common spatial patterns (CSP), convolution, electroencephalogram (EEG), neural networks, rapid serial visual presentation (RSVP), spatial filters.**

I. Introduction

EVENT-RELATED potentials (ERPs) are systematic voltage fluctuations caused by the postsynaptic neural activity of cortical pyramidal neurons that are time-locked to internal or external events [1]. ERPs are typically measured noninvasively using electrodes placed on the scalp and they are a direct high temporal resolution (<10 ms) measure of neural activity. Importantly, specific ERPs (voltage deflections at specific postevent time points) have been associated with a variety of perceptual and cognitive functions [2].

The specific ERPs evoked by a given individual performing a task can be relatively consistent in terms of both amplitude and latency [3]. The stability of specific ERPs has been leveraged in the application of brain–computer interfaces (BCIs) that use machine learning algorithms to detect specific ERP responses [4], [5]. For example, several groups have developed BCI spellers that are based on the detection of the P300 ERP [6] and the N200 ERP [7]. In spite of the relative stability of these ERPs, accurate and reliable detection of the specific neural response often requires averaging multiple responses. For instance, it is common that about ten individual P300 responses are averaged in BCI spellers to assure an optimal detection [8]. The need for averaging is largely due to noise in the electroencephalogram (EEG) that is not task-related and by the spatially diffuse distribution of the ERP across electrodes. Although averaging multiple ERP responses can increase the efficiency of detection, it impedes the information transfer rate of the BCI [9]. Therefore, there is an increasing emphasis on identifying methods that permit ERP detection using a single response.

New signal processing and machine learning methods have made efficient single-trial detection of ERPs possible and has extended the number of possible applications relying on ERP detection. One application that has received some attention in the recent BCI literature has been single-trial target detection in rapid serial visual presentation (RSVP) tasks [10], [11]. In the RSVP paradigm, a rapid sequence of images (e.g., 2–10 images per second) are presented to observers sequentially in the same location [12], [13]. The stream of images contains different types of visual stimuli that can be defined as targets or nontargets and depending on the specific task, the targets and nontargets elicit different ERPs. Several studies have used this paradigm during visual search [14]–[18], and face recognition tasks [19]. The strength of the RSVP paradigm is that the speed of the stimulus sequence combined with single-trial ERP detection increases the upper limit of potential information transfer rates in BCI applications. However, despite the possible use of single-trial ERP detection in RSVP tasks, efficient detection remains a

Manuscript received February 7, 2013; revised July 17, 2013 and October 11, 2013; accepted January 19, 2014. This work was supported by the Institute for Collaborative Biotechnologies under Contract W911NF-09-D-0001 through the U.S. Army Research Office.

H. Cecotti is with the Department of Psychological and Brain Sciences, Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA 93106 USA, and also with the School of Computing and Intelligent Systems, University of Ulster, Londonderry BT48 7JL, U.K. (e-mail: hub20xx@hotmail.com).

M. P. Eckstein and B. Giesbrecht are with the Department of Psychological and Brain Sciences, Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA 93106 USA (e-mail: eckstein@psych.ucsb.edu; giesbrecht@psych.ucsb.edu).

Digital Object Identifier 10.1109/TNNLS.2014.2302898

difficult problem and an active research area. For example, in a recent international machine learning workshop competition (MLSP) [20] using an RSVP task, the area under the receiver operating characteristic (ROC) curve (AUC) of the best participants reached about 0.82.

There are multiple potential signal processing approaches that may increase the efficiency of detecting ERPs using a single measured response. One approach that is the focus of this paper is spatial filtering. Spatial filtering refers to methods that change the EEG signal in a manner that enhances the relevant information contained in the signal. Spatial filtering plays an important role in EEG analysis due to the background noise of the EEG and because of the diffusion of the signal caused by the dura, skull, and scalp. Working with high quality signals is even of greater importance in single-trial analysis. The efficacy of spatial filtering in the detection process has been demonstrated in the P300 speller paradigm [21] and for motor imagery based BCIs [22]. What is unclear from these studies is the extent to which the improvements gained using spatial filtering are dependent on the subsequent classification approach. Indeed, Parra *et al.* [23] demonstrate the utility of linear analysis methods for discriminating between different events in single-trial, stimulus driven experimental paradigms using EEG and MEG. Other efficient strategies without spatial filtering have been proposed for EEG single-trial detection. Those methods include linear classifiers, such as Fisher's linear discriminant analysis, Bayesian linear discriminant analysis (BLDA) [24], and support vector machines (SVM) [8], [25].

The primary purpose of this paper was to investigate the effects of spatial filtering on the single-trial detection of ERPs recorded during an RSVP task. To investigate this issue, we used a convolutional neural network (CNN) with an embedded spatial filtering approach in which the filtering and classification are performed in an united way. Unlike previous CNN applications for the detection of ERP in the P300 speller [26], we propose a learning approach based on the maximization of AUC. Because it is unknown whether spatial filtering is required for obtaining optimal performance and whether the performance achieved by the embedded spatial filtering in the CNN could also be achieved with a different method, we also used two different state-of-the-art spatial filtering methods and three classifiers. The two spatial filtering methods were: xDAWN and common spatial pattern (CSP). The classifiers included both linear classifiers (BLDA and SVM) and a multilayer perceptron (MLP). Each combination of spatial filtering and classification approach were evaluated on EEG data from three separate RSVP experiments, using a total of 28 data sets. In addition, we also assessed the extent to which the single-trial classification performance was modulated by behavioral performance relative to ground truth.

This paper is organized as follows. First, we present the general rationale for the spatial filtering approach. Second, we present specific spatial filtering methods used in this paper. Third, we describe the classification methods and performance evaluation metrics. Fourth, we present the experimental methods. Finally, the results are presented and discussed in the last two sections.

II. Spatial Filtering

The purpose of spatial filtering is to enhance a particular subset of information that is contained in the EEG signal by creating virtual electrodes, or virtual sensors. We denote by channel the notion of virtual electrodes, which reduce the number of features. Channels represent a weighted combination of the electrode inputs. Here, we make two main assumptions about the data that influences the creation of the spatial filters. First, multiple sets of spatial filters can be created for a given data set, but because of the stability of the ERP to specific stimuli, we assume that the same set of spatial filters can be applied across the whole signal because the characteristics of the stimuli during the experiment are stable. Indeed, the latency and amplitude of the ERP may vary over time for a given task in relation to different experimental parameters, such as the target probability and the stimulus meaning [27], [28]. Second, it is assumed that the ERP waveform is spatially stationary. These assumptions allow us to apply uniformly a single set of spatial filters over time, the goal of which is to enhance the relevant information. A key aspect of these filters is that each of the sensors is weighted relative to the discriminative information it carries. Because of the creation of channels and the relative weighting, filtering reduces the number of input features for the classifiers because the number of channels (N_c) is usually inferior to the number of sensors (N_s), $N_c < N_s$.

We define a channel as a linear combination of the signals measured by the N_s sensors. A channel c is defined by N_s weights. At each time j the output value of a channel c is

$$c_j = \sum_{i=1}^{N_s} w_i I_{i,j} \tag{1}$$

where I is a 2-D signal, $0 \leq i < N_s$.

Without spatial filtering, a channel i_1 corresponds to a sensor i_1 (if $i = i_1$ then $w_i = 1$, $w_i = 0$ otherwise). The creation of virtual sensors results in a set of spatially independent vectors. The information from the different sensors is condensed in one scalar at a time j. The goal is to find an optimal set $w(k)_i$, $0 \leq i < N_s$, and $0 \leq k < N_c$.

A channel usually represents the effect of a spatial filter. Several approaches are presented in the literature for setting adaptive spatial filters, including statistical methods like independent component analysis (ICA) and CSP [22], [29]–[32]. We distinguish supervised spatial filtering method like CSP, and other methods like ICA that do not rely on the knowledge of the stimulus onsets. While ICA can be used for spatial filtering, its main purpose is blind source separation and not to discriminate classes. In this paper, we focus on efficient spatial filtering methods that require the knowledge of the stimulus onsets, where the goal is to enhance differences between two classes. In the following subsections, we consider two state-of-the-art methods: xDAWN and CSP, which are based on the Rayleigh quotients. For xDAWN, the goal is to maximize the signal-to–signal-plus noise ratio (SSNR) whereas for CSP, the goal is to maximize the ratio between the discriminative activity and the common activity, leading to optimal variances for the discrimination of two types of signals [22].

A. xDAWN

The xDAWN algorithm has been successfully applied in BCI for ERP detection and sensor selection in the P300 speller paradigm [21], [33]–[35]. It considers an algebraic model of the recorded signals X of size $N_t \times N_s$, where N_t is the number of temporal samples in the recorded EEG signals. X is composed of three terms: 1) the responses on targets ($D_1 A_1$); 2) a response common to all stimuli, i.e., targets and nontargets confound ($D_2 A_2$); and 3) the residual noise (H)

$$X = D_1 A_1 + D_2 A_2 + H \tag{2}$$

where D_1 and D_2 are two real Toeplitz matrices of size $N_t \times N_1$ and $N_t \times N_2$, respectively. D_1 has its first column elements set to zero except for those that correspond to a target onset, which are represented with a value equal to one. For D_2, its first column elements are set to zero except for those that correspond to stimulus onset. N_1 and N_2 are the number of sampling points representing the target (the ERP response on the target) and superimposed evoked potentials, respectively. A_1 and A_2 are matrices of size $N_1 \times N_s$ and $N_2 \times N_s$, respectively. H is a real matrix of size $N_t \times N_s$.

The goal of applying spatial filters U is to enhance the SSNR of the responses corresponding to the presentation of a target ($D_1 A_1 U$), where N_f is the number of spatial filters

$$XU = D_1 A_1 U + D_2 A_2 U + HU. \tag{3}$$

We define the SSNR in relation to the spatial filters by

$$\mathrm{SSNR}(U) = \frac{Tr(U^T \hat{A}_1^T D_1^T D_1 \hat{A}_1 U)}{Tr(U^T X^T X U)} \tag{4}$$

where $Tr(.)$ is the trace of the matrix, and $\hat{A}_1$ corresponds to the least mean square estimation of A_1

$$\hat{A} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix} \tag{5}$$

$$= ([D_1; D_2]^T [D_1; D_2])^{-1} [D_1; D_2]^T X \tag{6}$$

where $[D_1; D_2]$ is a matrix of size $N_t \times (N_1 + N_2)$ obtained by concatenation of D_1 and D_2.

Spatial filters are obtained through the Rayleigh quotient after two QR decompositions and a singular value decomposition by maximizing the SSNR [36]. More details about the computational method can be found in [21]

$$\hat{U} = \underset{U}{\operatorname{argmax}}\ \mathrm{SSNR}(U). \tag{7}$$

B. Common Spatial Pattern

CSP is one of the most used spatial filtering method in discriminating different classes in motor imagery based BCIs, where the task is to classify two different states of brain activity, e.g., imagery of the movement of the left or the right hand [22], [37]–[40]. Although this method is mainly applied for motor imagery, we consider here this method for ERPs. The CSP feature extraction can be estimated and interpreted using the framework of Rayleigh coefficient maximization.

First, two covariance matrices Σ_0 and Σ_1 are calculated for the two classes (target and nontarget)

$$\Sigma_i = \sum_{j \in C_i} \frac{E_j \cdot E_j^T}{Tr(E_j \cdot E_j^T)} \tag{8}$$

where $E_j \in \mathbb{R}^{N_s \times N_1}$ denotes an EEG data matrix of the jth trial.

The CSP method aims at finding a spatial filter w that maximizes the difference in the average band power of the filtered signal while keeping the sum constant

$$D = w \Sigma_0 w^T \tag{9}$$

$$I - D = w \Sigma_1 w^T \tag{10}$$

where D is a diagonal matrix and I is the identity matrix. Then, we shall construct a matrix U which is composed of the first and last components of u, which correspond to the first and last ordered eigenvalues.

The CSP matrix is extracted thanks to the Rayleigh coefficient maximization by solving a generalized eigenvalue problem

$$\hat{U} = \underset{U}{\operatorname{argmax}}\ \frac{U^T R_0 U}{U^T R_1 U} \tag{11}$$

where R_0 and R_1 represent the discriminative and common activity, respectively, and are defined as follows:

$$R_0 = \Sigma_0 - \Sigma_1 \tag{12}$$

$$R_1 = \Sigma_0 + \Sigma_1. \tag{13}$$

The discriminative activity corresponds to the differences between targets and nontargets whereas the common activity corresponds to what is common to targets and nontargets. The spatial filters $\mathbf{u_c}$ are based on the eigenvectors from both ends of the eigenvalue spectrum. Like for $\mathrm{MLP_{cnn}}$, we consider four spatial filters $N_c = 4$. Thus, the set of spatial filters $\hat{U} \in \mathbb{R}^{N_s \times N_c}$ is

$$\hat{U} = [\mathbf{u_0}, \mathbf{u_1}, \mathbf{u_{N_s-2}}, \mathbf{u_{N_s-1}}]. \tag{14}$$

III. Methods

We tested three classifiers using four different spatial filtering approaches. The first method was based on a CNN ($\mathrm{MLP_{cnn}}$), and therefore includes the classifier. In this method, spatial filtering was a part of the neural architecture. The set of spatial filters is tuned in relation to their discriminant power once they are combined for classification. We denote by $\mathrm{MLP_{csp}}$and $\mathrm{MLP_{xdawn}}$, the method with CSP and xDAWN as spatial filtering, respectively. With each spatial filtering approach, we used three different classifers, MLP, a BLDA classifier [24], [41], and a linear SVM [42], [43]. Finally, each of the classifiers with filtering was compared with the same classifiers without filtering. The factorial combination of classifier $\times$ spatial filtering approaches resulted in 12 separate methods.

A. Convolutional Neural Network

CNNs are efficient for handwriting character recognition [44], [45], vision [46], and the classification of EEG signals [26], [47], [48]. It allows extracting some particular features through its layers that are directly learned in relation to the problem to solve. This type of neural network has many advantages when the input data contains an inner structure like for images (2-D) and some signals (2-D: times × space), when the features are difficult to model in an analytical way [49].

The neural network for the detection of ERP is composed of four layers. Each layer is composed of at least one map. In the first and second hidden layers each map represents the signal after spatial filtering. The neural architecture is described as follows.

1) The input layer (L_0): $I(i, j)$ with $1 \leq i \leq N_s$ and $1 \leq j \leq N_1$. In the experiment, $N_s = 32$ and $N_1 = 26$. This layer corresponds to the EEG signal after temporal filtering and downsampling.
2) The first hidden layer (L_1): It is composed of N_c maps. In the experiment, we set $N_c = 4$ as it was used in [33]. We define $L_1 M_m$, as the map number m. Each map of L_1 has the size N_1. This layer corresponds to N_c channels. Each map corresponds to a projection of the EEG signal in 1-D (time). As the weights of each modeled neuron in a map are shared, the transformation from L_0 to L_1 is the application of a convolution filter. As the convolution is applied only across values in the space domain, the filter is equivalent to a spatial filter. In L_1, the EEG signal is represented as N_c vectors of size N_1.
3) The second hidden layer (L_2): It is composed of one map of 40 neurons. This map is fully connected to the different maps of L_1. The number of neurons was chosen in relation to previous tests on other database of brain responses [26], [47].
4) The output layer (L_3): This layer has only one map of M neurons, $M = 2$ for target and nontarget. This layer is fully connected to L_2. The first and second neuron have the expected value 1 and 0, respectively, when the input corresponds to a target, and 0 and 1 when it is not a target.

We define the value of a neuron in the layer l, in the map m at the position j by $x_{(l,m,j)}$, or $x_{(l,j)}$ when there is only one map in the layer. Similarly, we define $\sigma_{(l,m,j)}$ as the scalar product between a set of input neurons and the weight connection between these neurons and the neuron number j in the map m in the layer l

$$x_{(l,m,j)} = f(\sigma_{(l,m,j)}) \tag{15}$$

where f is a sigmoid function (hyperbolic tangent for L_1, logistic function for L_2 and L_3) [50].

We define $\sigma_{(l,m,j)}$ for the two hidden layers and the output layer. It is worth noting that L_1, L_2, and L_3 can be considered as an MLP where L_1 is the input layer, L_2 is the hidden layer, and L_3 is the output layer. For L_1, each neuron of one map shares the same set of weights.

1) For L_1

$$\sigma_{(1,m,j)} = wt_{(1,m,0)} + \sum_{i=0}^{i<N_s} I_{i,j} \cdot w_{(1,m,i)} \tag{16}$$

where $wt_{(1,0,j)}$ is a threshold. A set of weights $w_{(1,m,i)}$ with m fixed, $0 \leq i < N_r$ corresponds to a spatial filter (the convolution). In this layer, there are $N_s + 1$ weights for each map.

2) For L_2

$$\sigma_{(2,j)} = wt_{(2,0,j)} + \sum_{i=0}^{i<N_c} \sum_{k=0}^{k<N_1} x_{(1,i,k)} \cdot w_{(2,i,k)} \tag{17}$$

where $wt_{(2,0,j)}$ is a threshold. L_2 is fully connected to L_1. In this layer, each neuron has $N_c \cdot N_s + 1$ input weights.

3) For L_3

$$\sigma_{(3,j)} = wt_{(3,0,j)} + \sum_{i=0}^{i<40} x_{(2,i)} \cdot w_{(3,i)} \tag{18}$$

where $wt_{(3,0,j)}$ is a threshold. L_3 is fully connected to L_2.

The learning algorithm for tuning the weights and thresholds of the network uses the backpropagation [51], by maximizing the AUC of the validation database. In Section V, we will also compare learning based on the minimization of the mean square error (MSE) between the neurons in L_3 and their expected values. At the initialization of the network, the weights and the thresholds of each neuron are initialized randomly with a standard distribution around $\pm 1/N_{\text{input}}$ where N_{input} is the number of input links for each neuron. For training, the learning parameter was set to $\lambda = 0.3$. This model was implemented in C++ without any special hardware optimization.

B. Classifiers

As the spatial filtering step and the classification are united in MLP_{cnn}, we have considered an MLP for the classification when we consider spatial filters based on CSP or xDAWN to stay consistent with the classification method across the different spatial filtering approaches. The MLP possesses therefore the same parameters as the last three layers of the CNN, i.e., the number of input is $N_c \cdot N_1$, the number of neurons in the hidden layer is 40, and the output layer contains two neurons. For the evaluation of the classifier, we provide the results obtained after a five fold cross validation. Four blocks were considered for training the classifier and the spatial filters. The remaining block was used for testing the classifier. We considered 12.5% of the blocks dedicated for training as a validation database to determine when training should be stopped. As the classifier contains two outputs, we define the confidence value of the classifier as

$$y_{\text{out}} = \frac{y_0 + (1 - y_1)}{2} \tag{19}$$

where y_0 and y_1 are the states of the neurons in the output layer. For the CNN, we have $y_0 = x_{(3,0)}$ and $y_1 = x_{(3,1)}$.

TABLE I
NUMBER OF FREE PARAMETERS FOR EACH METHOD DURING TRAINING

Method	# inputs	# free parameters
MLP_{cnn} :	104	4414
MLP_{xdawn} :	104	4282
MLP_{csp} :	104	4282
$\text{MLP}_{\emptyset}$:	832	33402

C. Complexity

Table I presents the number of inputs and the number of free parameters in the neural network. MLP_{cnn}, $\text{MLP}_{\text{xdawn}}$, and MLP_{csp} have the same number of inputs as we consider the same number of sampling points and channels ($N_c \cdot N_1 = 4*26 = 104$). $\text{MLP}_{\text{xdawn}}$ and MLP_{csp} have the same number of free parameters as only the spatial filtering method changes. The only difference between MLP_{cnn} and ($\text{MLP}_{\text{xdawn}}$, MLP_{csp}) corresponds to the convolutional layer ($132 = (N_s + 1) \cdot N_c = (32 + 1) * 4$). The most complex network is $\text{MLP}_{\emptyset}$with 33402 connections and thresholds.

D. Performance Measures

The classifiers are then evaluated through ROC graphs based on the true positive rate (TPR) and false positive rate (FPR). We define the TPR and FPR as

$$\text{True positive rate (TPR)} = \frac{\text{TP}}{P} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{20}$$

$$\text{False positive rate (FPR)} = \frac{\text{FP}}{N} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{21}$$

where TP, FP, TN, and FN are the number of true positive, false positive, true negative, and false negative, respectively. ROC curves allow analyzing and visualizing the performance of classifiers. As the classifier output produces a confidence measure, it is possible to generate ROC curves and computer its AUC as described in [52]. In the following parts, we consider nonparametric ROC curves based on 100 points.

For the behavioral performance, which was a binary response, we consider the AUC as the normal cumulative distribution function of $d'/\sqrt{2}$ where d' is the sensitivity index $d' = Z(\text{TPR}) - \text{Z(FPR)}$ and $Z(p)$, $p \in [0, 1]$, is the inverse of the cumulative Gaussian distribution.

IV. EXPERIMENTS

The 12 different methods were evaluated using three RSVP tasks that were performed by subjects recruited through the University of California, Santa Barbara (UCSB) online subject recruitment system. All procedures were approved by the UCSB Human Subjects Committee.

A. Experiment 1

Grayscale images (256×256 pixel) of faces (target) and cars (nontarget) were presented as stimuli to the observers who performed the behavioral task of identifying the correct label of the image (face/car). Fig. 1 shows examples of the visual stimuli with and without noise (participants only saw the versions with noise). These images were taken from the Max Planck Institute for Biological Cybernetics face database [53]. Each image was presented for 500 ms and immediately replaced by the subsequent image, resulting in a presentation rate of 2 Hz (visual angle ≈ 4.57°). In each session, target probability was 0.10. The target-distractor sequences were generated in blocks of 2 min, keeping a relative constant target probability over time. Eight healthy subjects participated in the experiment (mean = 23.5, sd = 8.38, three females). They were instructed to respond to the presence of a face (target) as quickly and accurately as possible by hitting the enter key on a standard keyboard.

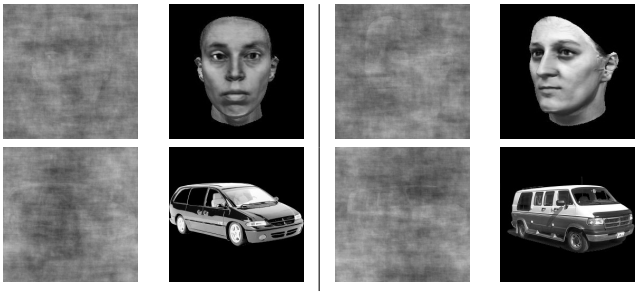


Fig. 1. Samples of visual targets (faces, top) and nontargets (cars, bottom) with their corresponding model.

TABLE II
NUMBER OF PATTERNS FOR EACH CLASS CORRESPONDING TO A CORRECT BEHAVIORAL RESPONSE

Subject	Faces	Cars
1	994	10556
2	1086	9952
3	624	10692
4	1057	4599
5	890	9597
6	998	10280
7	1087	3278
8	1037	6585

For the binary classification of target versus nontarget, we consider two different problems in relation to the behavioral performance.

1) The binary classification of target versus nontarget independent of the behavioral performance (target versus nontarget). It is the classification of the ERPs corresponding to every target and nontarget. The EEG database contains 1200 patterns representing the target (faces) and 10800 for the nontarget (cars), for each subject.
2) The binary classification of target versus nontarget dependent of the behavioral performance (target hit versus nontarget hit). In this case, we consider only the patterns that correspond to a correct behavioral response, i.e., the subject presses a button when there is a target and does not press a button when there is no target. The analysis of a behavioral response is considered between 0.2 and 1 s after the presentation of a stimulus. The number of patterns for both classes and for each subject is detailed in Table II.

Fig. 2. Examples of visual stimuli [(left) target versus (right) nontarget].

B. Experiment 2

Visual stimuli consisted of 900 color images (683×384 pixel). These images were taken from Insurgency: Modern Infantry Combat (Insurgency Team), a total conversion modification of the video game Half-Life 2 (Valve Corporation). The realistic images were separated into target scenes that contained a person (300 images) and nontarget scenes that did not contain a person (600 images). Several examples of the images that were presented during the experiments are shown in Fig. 2. The images were centered on the screen (visual angle $\approx 26°$) and presented for 100 ms with no interval between subsequent images. The RSVP task was separated into different blocks of 5 s. Participants were instructed to hold their blinks to avoid eye blink artifacts in the EEG data. They had to count the number of times they saw a target (image with a person) and report the number at the end of the block. The probability of a target was set 0.10. Each block contained ten different images, one of them being a target and the experiment consisted of 5000 images. Ten healthy subjects (mean = 19.5, sd = 1, five females) participated in the experiment. Each session contained 5000 images. Therefore, the database contains 1000 patterns representing the target (someone in the scene) and 9000 for the nontarget class (nobody in the scene), for each subject.

C. Experiment 3

The third experiment was identical to the second experiment, except that stimuli were presented for 200 ms. The same subjects of Experiment 2 participated in the RSVP task in two sessions. Each session contained 2000 images. This database contains 400 patterns representing the target (someone in the scene) and 3600 for the nontarget class (nobody in the scene), for each subject.

D. Signal Acquisition

The EEG signal was recorded from 32 Ag/AgCl sintered sensors mounted in an elastic headcap (Biosemi ActiveTwo). The 32 electrodes were placed according to a subsampled version of the 10–10 system [54]. The horizontal and vertical electrooculograms were recorded from sensors placed 1 cm lateral to the external canthi (left and right) and above and below each eye, respectively. The data were sampled at 512 Hz and referenced offline to the signal recorded from the mastoids.

E. Signal Preprocessing

The signal was first bandpassed filtered (Butterworth filter of order 4) with cutoff frequencies at 1 and 10.66 Hz. Then, the signal was downsampled to obtain a signal at a sampling frequency equivalent to 32 Hz. This sampling frequency was used by the winning team of the MLSP competition 2010 [55]. The resulting signal used for target detection included amplitude values (in microvolts) between 0 and 812.5 ms after the start of a visual stimulus (26 sampling points, $N_1 = 26$), during which we assumed that the targets should evoke enhanced ERP responses (e.g., P300).

V. Results

A. Experiment 1

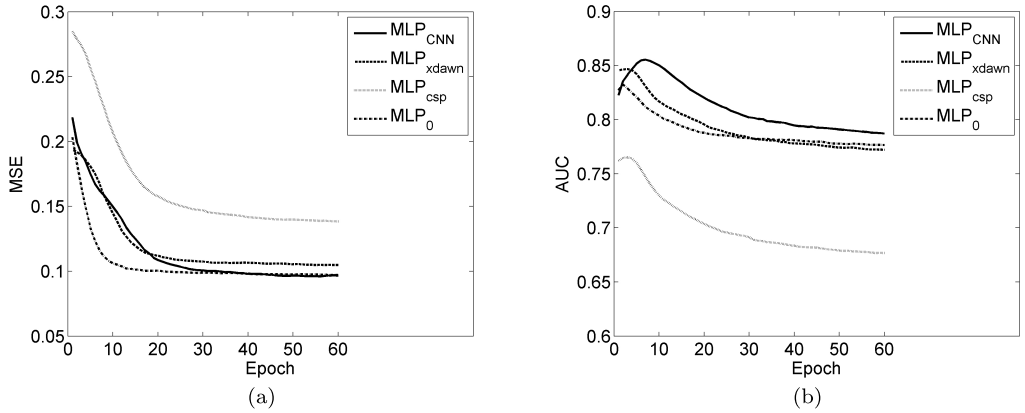
The performance for target detection are presented in two different ways. First, the methods are assessed in their ability to detect the type of stimulus (car or face). In this case, the ground truth is objective, based on the presentation of cars and faces (independent on behavioral performance). Second, the methods are assessed in their ability to detect the type of stimulus based on the behavioral decision of the observer. For this condition, the ground truth is subjective, based on the subjects identification of cars and faces (dependent on behavioral performance).

1) Performance Classifying Stimulus: The AUC for each subject, the mean and standard deviation (SD) across subjects are presented in Table III. The last row of the table shows the subjects' behavioral performance. The best mean accuracy was achieved with the CNN (MLP_{cnn}) with a mean AUC of 0.861 ± 0.073. There was a significant difference across the 12 methods (Friedman's test, $p < 10e-5$). After posthoc analysis with a false discovery rate correction, the best preprocessing method was CNN, followed by xDAWN and the absence of preprocessing method (there was no difference between xDAWN and the absence of spatial filters), and CSP (Wilcoxon sign rank test $p < 0.01$). For the classification step, MLP was better than BLDA ($p < 10e-5$) and SVM ($p < 10e-5$), and there was no difference between BLDA and SVM. The evolution of the MSE and the AUC across the different epoch during the neural network training on the validation database is presented in Fig. 3. A pairwise t-test indicated that the performance based on the maximization of the AUC was superior to the minimization of MSE ($t_{31} = 11.803$, $p < 10e-5$). The mean AUC performance based on the AUC maximization and MSE minimization is 0.825 ± 0.083 and 0.764 ± 0.105, respectively. The training step converges after 4.5 ± 2.7 iterations for the maximization of the AUC whereas it requires 42.7 ± 7.7 iterations for MSE minimization.

The resulting spatial filters of a representative subject (Subject 2) are depicted as topographic maps of the scalp in Fig. 4. The different gray values correspond to the weight values of the different spatial filters. For MLP_{cnn}, each map corresponds to the values of the weights that are used between

TABLE III
EXPERIMENT 1: AUC FOR EACH SUBJECT AND EACH METHOD (TARGET VERSUS NONTARGET). FOR EACH SUBJECT, THE BEST RESULT IS DISPLAYED IN BOLD CHARACTERS

	Subject									
Method	1	2	3	4	5	6	7	8	Mean	SD
MLP_{cnn}	**0.922**	**0.947**	**0.867**	0.716	**0.819**	**0.937**	**0.810**	**0.868**	**0.861**	0.73
MLP_{xdawn}	0.910	0.917	0.851	**0.744**	0.786	0.913	0.789	0.842	0.844	0.62
MLP_{csp}	0.831	0.888	0.782	0.642	0.681	0.832	0.725	0.732	0.764	0.78
$\text{MLP}_{\emptyset}$	0.892	0.929	0.829	0.701	0.758	0.927	0.775	0.831	0.830	0.77
BLDA_{cnn}	0.902	0.917	0.858	0.737	0.778	0.919	0.781	0.834	0.841	0.66
BLDA_{xdawn}	0.906	0.910	0.853	**0.744**	0.783	0.911	0.778	0.834	0.840	0.62
BLDA_{csp}	0.826	0.876	0.742	0.606	0.641	0.818	0.704	0.696	0.739	0.89
$\text{BLDA}_{\emptyset}$	0.857	0.913	0.738	0.653	0.670	0.885	0.744	0.792	0.782	0.91
SVM_{cnn}	0.913	0.937	0.858	0.564	0.723	0.930	0.682	0.837	0.806	0.127
SVM_{xdawn}	0.911	0.910	0.854	0.532	0.506	0.912	0.532	0.616	0.722	0.179
SVM_{csp}	0.801	0.879	0.599	0.515	0.548	0.608	0.519	0.518	0.624	0.131
$\text{SVM}_{\emptyset}$	0.912	0.927	0.845	0.550	0.740	0.922	0.734	0.832	0.808	0.121
Behavioral	0.982	0.973	0.954	0.759	0.908	0.968	0.714	0.835	0.886	0.105



(a) (b)

Fig. 3. (a) MSE and (b) AUC evolution across epochs during the neural network training on the validation database (Experiment 1—stimulus classification).

L_0 and L_1. The topographic maps show the differences between the spatial filters obtained with CNN (MLP_{cnn}), xDAWN, and CSP. It is important to stress that these plots represent the weights on each electrode of the spatial filters and do not represent the spatial distribution of the ERP across the scalp. For the spatial filters obtained with the CNN, the order of the filters has no meaning, contrary to xDAWN and CSP as described in Sections II-A and II-B. The first spatial filter of xDAWN corresponds to the maximization of the SSNR and represents the best filter for maximizing the SSNR.

2) Performance Predicting Single-Trial Behavioral Decisions: The AUC for each subject, the mean and SD across subjects are presented in Table IV. The best mean accuracy was achieved with $\text{MLP}_{\text{cnn}}$with an AUC of 0.932 ± 0.034. There was a significant difference across the 12 methods (Friedman's test, $p < 10e-5$). After posthoc analysis with a false discovery rate correction, we find the same pattern of performance than for the classification of the stimuli, the best preprocessing method was CNN, followed by xDAWN and the absence of preprocessing method (no difference between xDAWN and the absence of spatial filters), and CSP (Wilcoxon sign rank test $p < 0.01$). For the classification step, MLP was better than BLDA ($p < 10e-5$), SVM was also better than BLDA ($p < 0.005$), and there was no difference between MLP and SVM. Like for performance classifying stimulus, the evolution of the MSE and the AUC during the training of the different neural networks for predicting single-trial behavioral decisions is given in Fig. 5. The performance based on the maximization of the AUC was superior to the minimization of MSE ($t_{31} = 6.939$, $p < 10e-5$). The mean AUC performance based of the AUC maximization and MSE minimization is 0.895 ± 0.058 and 0.864 ± 0.075, respectively. The training step converges after 7.8 ± 6.4 iterations for the maximization of the AUC whereas it requires 33.54 ± 9.3 iterations for MSE minimization.

3) Differences Between Classifying Stimulus and Behavioral Decisions: To compare the effect of the type of ground truth (subjective based on behavioral performance and objective based on the stimulus types) and spatial filtering, a repeated-measures two-way ANOVA was performed and indicated an effect on both the type of ground truth ($F(1,7) = 14.63$, $p < 10\text{e}-2$) and the spatial filters ($F(3,21) = 35.93$, $p < 10\text{e}-7$) but there was no interaction between them ($F(3,21) = 0.73$, $p = 0.54$). Similarly, a repeated-measures two-way ANOVA was performed to compare the effect of the type of ground truth and classifier. It revealed an effect on the type of ground truth ($F(1,7) = 14.63$, $p < 10\text{e}-2$), the spatial filters ($F(2,14) = 10.97$, $p < 10\text{e}-2$), and on the interaction ($F(2,14) = 9.84$, $p < 10\text{e}-2$). These results show that the ground truth for the performance estimation of

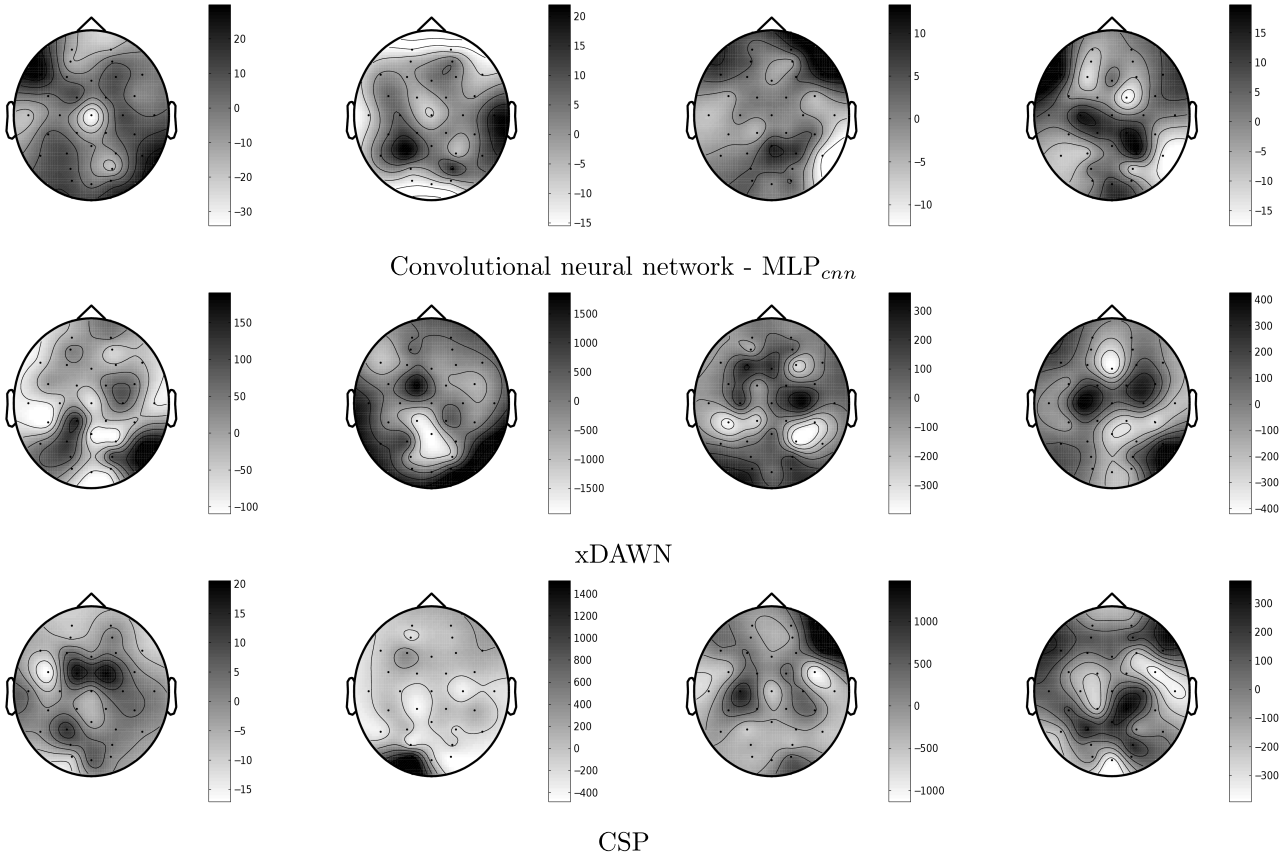


Fig. 4. Example of spatial filters obtained with MLP_{cnn}, $\text{MLP}_{\text{xdawn}}$, and MLP_{csp} for the subject 2 (all units are arbitrary).

TABLE IV
EXPERIMENT 1: AUC FOR EACH SUBJECT AND EACH METHOD (TARGET HIT VERSUS NONTARGET HIT). FOR EACH SUBJECT, THE BEST RESULT IS DISPLAYED IN BOLD CHARACTERS

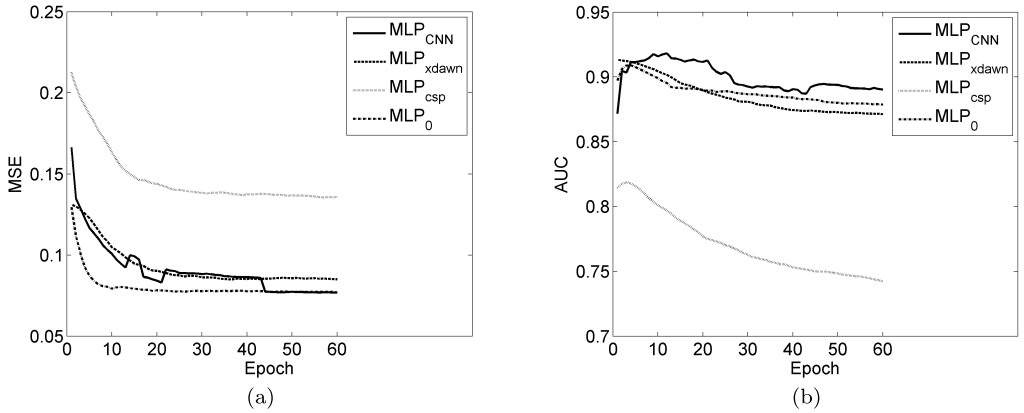
	Subject									
Method	1	2	3	4	5	6	7	8	Mean	SD
MLP_{cnn}	**0.946**	**0.961**	0.939	0.924	**0.894**	0.953	**0.902**	**0.939**	**0.932**	0.22
MLP_{xdawn}	0.942	0.935	0.941	0.866	0.855	0.944	0.883	0.913	0.910	0.34
MLP_{csp}	0.865	0.900	0.781	0.761	0.749	0.862	0.820	0.835	0.822	0.51
$\text{MLP}_{\emptyset}$	0.936	0.947	0.943	0.934	0.845	0.953	0.868	0.913	0.917	0.37
BLDA_{cnn}	0.937	0.938	0.943	**0.958**	0.852	0.948	0.878	0.903	0.920	0.36
BLDA_{xdawn}	0.938	0.931	0.939	0.868	0.853	0.949	0.883	0.907	0.908	0.34
BLDA_{csp}	0.843	0.890	0.566	0.752	0.695	0.848	0.799	0.819	0.776	0.97
$\text{BLDA}_{\emptyset}$	0.882	0.933	0.777	0.805	0.737	0.913	0.854	0.880	0.848	0.64
SVM_{cnn}	**0.946**	0.956	**0.948**	0.956	0.875	**0.955**	0.892	0.916	0.931	0.30
SVM_{xdawn}	0.944	0.931	0.946	0.868	0.842	0.950	0.884	0.907	0.909	0.38
SVM_{csp}	0.845	0.894	0.530	0.614	0.563	0.854	0.810	0.842	0.744	0.139
$\text{SVM}_{\emptyset}$	0.942	0.948	0.937	0.854	0.843	0.954	0.883	0.911	0.909	0.41

single-trial detection has a significant impact. In addition, it revealed that the difference obtained between the two ground truths is due to the classifier and not the spatial filtering method.

B. Experiment 2

The methods are assessed in their ability to detect the type of stimulus (images containing a person versus images containing nobody). There was a significant difference across the 12 methods (Friedman's test, $p < 10e-5$). After posthoc analysis with a false discovery rate correction, the best preprocessing method was xDAWN, followed by CNN, the absence of preprocessing method, and CSP (Wilcoxon sign rank test $p<0.01$). For the classification step, MLP was better than BLDA (p=0.0115), and there was no difference between MLP and SVM. The AUC for each subject, the mean and SD across subjects are presented in Table V. The best mean accuracy was achieved with $\text{BLDA}_{\text{xdawn}}$ with an AUC of 0.869 ± 0.051.

The evolution of the MSE and the AUC across the different epoch during the neural network training is shown in Fig. 6. A pairwise t-test indicated that the performance based on the maximization of the AUC was superior to the minimization



(a) (b)

Fig. 5. (a) MSE and (b) AUC evolution across epochs during the neural network training on the validation database (Experiment 1—behavioral response classification).

TABLE V
EXPERIMENT 2: AUC FOR EACH SUBJECT AND EACH METHOD (TARGET VERSUS NONTARGET). FOR EACH SUBJECT, THE BEST RESULT IS DISPLAYED IN BOLD CHARACTERS

	Subject											
Method	1	2	3	4	5	6	7	8	9	10	Mean	SD
MLP_{cnn}	0.852	0.759	0.898	0.846	0.835	**0.929**	0.897	0.762	0.756	0.915	0.845	0.63
MLP_{xdawn}	0.873	**0.817**	0.919	0.846	0.852	0.928	0.908	0.798	0.774	0.931	0.865	0.54
MLP_{csp}	0.751	0.653	0.840	0.607	0.678	0.812	0.802	0.585	0.674	0.795	0.720	0.87
$\text{MLP}_{\emptyset}$	0.860	0.728	0.895	0.831	0.595	0.904	0.900	0.728	0.741	0.913	0.809	0.101
BLDA_{cnn}	0.863	0.756	0.903	0.847	0.844	0.926	0.903	0.777	0.767	0.917	0.850	0.61
BLDA_{xdawn}	**0.877**	0.814	0.919	0.848	**0.859**	0.927	**0.912**	0.804	**0.791**	**0.938**	**0.869**	0.51
BLDA_{csp}	0.700	0.560	0.834	0.508	0.647	0.799	0.709	0.523	0.477	0.788	0.654	0.125
$\text{BLDA}_{\emptyset}$	0.649	0.562	0.851	0.677	0.611	0.819	0.740	0.589	0.552	0.892	0.694	0.118
SVM_{cnn}	0.857	0.771	0.900	**0.850**	0.841	0.922	0.903	0.757	0.753	0.917	0.847	0.63
SVM_{xdawn}	0.871	0.814	**0.920**	**0.850**	0.854	0.925	0.911	0.574	0.492	**938**	0.815	0.147
SVM_{csp}	0.518	0.548	0.630	0.479	0.571	0.657	0.530	0.507	0.507	0.512	0.546	0.54
$\text{SVM}_{\emptyset}$	0.864	0.785	0.873	0.837	0.813	0.910	0.892	0.759	0.748	0.923	0.840	0.59

of MSE ($t_{39} = 14.790$, $p < 10e-5$). The mean AUC performance based of the AUC maximization and MSE minimization is 0.791 ± 0.111 and 0.726 ± 0.119, respectively. The training step converges after 4.0 ± 2.7 iterations for the maximization of the AUC whereas it requires 41.7 ± 11.2 iterations for MSE minimization.

C. Experiment 3

As in Experiment 2, the analytical techniques were assessed in their ability to detect the type of stimulus (images containing a person versus images containing nobody). By considering the AUC, there was a significant difference across the 12 methods (Friedman's test, $p < 10e-5$). After posthoc analysis with a false discovery rate correction, the best preprocessing method was xDAWN, followed by CNN, the absence of preprocessing method, and CSP (Wilcoxon sign rank test $p < 0.01$). For the classifier only, MLP was better than BLDA ($p = 0.0013$), and there was no difference between MLP and SVM. The AUC for each subject, the mean and SD across subjects are presented in Table VI. The best mean accuracy was achieved with $\text{BLDA}_{\text{xdawn}}$ with an AUC of 0.854 ± 0.039.

The evolution of the MSEand the AUC across the different epoch during the neural network training on the validation database is presented in Fig. 7. A pairwise t-test indicated that there was no difference in performance between the maximization of the AUC and the minimization of MSE. The mean AUC performance based of the AUC maximization and MSE minimization is 0.773 ± 0.075 and 0.759 ± 0.086, respectively. The training step converges after 6.7 ± 4.6 iterations for the maximization of the AUC whereas it requires 37.5 ± 12.4 iterations for MSE minimization.

D. Performance Across Database

By considering the results obtained from the three experiments for the classification of the stimulus, we observed a significant difference across the 12 methods (Friedman's test, $p < 10e-5$). After Wilcoxon sign rank tests, CNN was the best preprocessing step, followed successively by xDAWN, the absence of spatial filtering, and CSP ($p < 10e-4$). For classifiers, MLP was better than both SVM ($p < 10e-5$) and BLDA ($p < 10e-5$), and there was no difference between SVM and BLDA.

VI. DISCUSSION

In this paper, we addressed three main issues. First, we investigated the efficacy of a CNN based on the maximization of the AUC for single-trial ERP detection in three RSVP tasks. This method embeds both the spatial filtering and the classification steps and outperforms the other methods in some conditions (BLDA and SVM with or without spatial filters, and MLP). Second, we have shown that spatial filtering is not

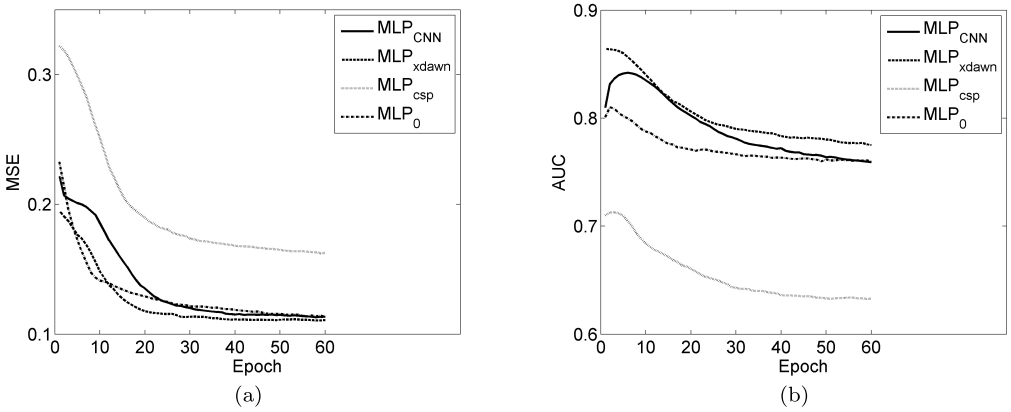


Fig. 6. (a) MSE and (b) AUC evolution across epochs during the neural network training on the validation database (Experiment 2).

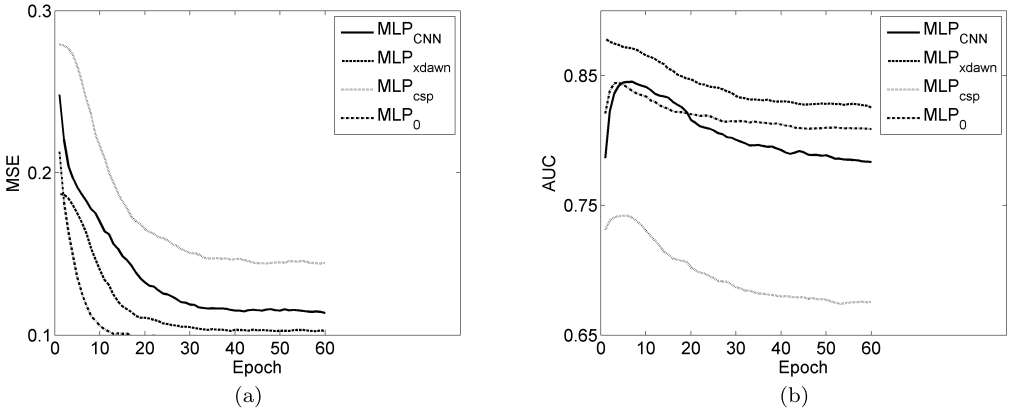


Fig. 7. (a) MSE and (b) AUC evolution across epochs during the neural network training on the validation database (Experiment 3).

TABLE VI
EXPERIMENT 3: AUC FOR EACH SUBJECT AND EACH METHOD (TARGET VERSUS NONTARGET).
FOR EACH SUBJECT, THE BEST RESULT IS DISPLAYED IN BOLD CHARACTERS

	Subject											
Method	1	2	3	4	5	6	7	8	9	10	Mean	SD
MLP_{cnn}	0.828	0.801	0.762	0.774	0.807	0.832	0.863	0.831	0.736	0.930	0.816	0.52
MLP_{xdawn}	0.861	**0.847**	**0.792**	0.830	0.838	**0.867**	0.853	0.846	0.774	0.946	0.845	0.44
MLP_{csp}	0.753	0.656	0.665	0.700	0.722	0.721	0.784	0.776	0.642	0.858	0.728	0.64
$\text{MLP}_{\emptyset}$	0.832	0.815	0.743	0.788	0.802	0.832	0.832	0.829	0.698	0.931	0.810	0.58
BLDA_{cnn}	0.847	0.804	0.772	0.802	0.792	0.862	0.871	0.831	0.730	0.926	0.824	0.53
BLDA_{xdawn}	**0.878**	0.843	0.807	**0.842**	0.840	0.865	0.871	**0.858**	**0.794**	0.943	**0.854**	0.39
BLDA_{csp}	0.622	0.622	0.646	0.556	0.619	0.522	0.697	0.736	0.563	0.833	0.642	0.88
$\text{BLDA}_{\emptyset}$	0.621	0.592	0.656	0.570	0.610	0.643	0.736	0.657	0.544	0.879	0.651	0.91
SVM_{cnn}	0.835	0.808	0.735	0.792	0.808	0.841	0.854	0.836	0.745	0.939	0.819	0.55
SVM_{xdawn}	0.860	0.835	0.483	0.833	**0.848**	0.856	0.873	0.856	0.623	**0.949**	0.802	0.132
SVM_{csp}	0.518	0.545	0.553	0.528	0.475	0.537	0.768	0.516	0.526	0.849	0.581	0.117
$\text{SVM}_{\emptyset}$	0.802	0.798	0.739	0.801	0.810	0.830	0.835	0.824	0.719	0.927	0.809	0.53

a necessary component of the system but it can improve the overall performance. Finally, we have shown that the estimation of the ground truth based on behavioral performance has a significant effect on single-trial detection performance. The implications of these results are described thereafter.

A CNN based on the maximization of the AUC is an efficient approach for the classification of ERPs during RSVP task because it does not require prior knowledge about the type of spatial filters to consider and the association between the classifier and the spatial filters. Thanks to the maximization of the AUC [56], this approach avoids pitfalls due to the unbalanced distribution of the classes in the data set and allows to maximize the AUC on the test data sets. Despite the advantages of this method, the choice of the number neurons, and the number of spatial filters, like for the other methods, requires *a priori* information from separate experiments or previously published data. Nevertheless, if this information is available to the experimenter, as it was here, then the results observed here indicate that the CNN approach can be effective at single-trial classification of ERPs in a difficult task.

The present observation that the efficiency of single-trial detection was dependent on both the spatial filtering approach and classifier demonstrates the importance of careful consideration of all components of system architecture is critical. Each

component of the system requires special attention to achieve optimal classifier performance given the quality of the data (the number of samples, the type of noise), and spatial filtering is no exception. One challenge is to design a robust strategy that keeps a coherent architecture between the different processing steps and is not solely focused on one stage (e.g., classification [57]). In several previous EEG detection studies [21], [22], the focus was toward feature extraction methods like spatial filtering and classifiers that were often linear (e.g., LDA). The conclusion based on these studies was that spatial filtering was required to achieve the best performance. However, we observed that an MLP or an SVM does not require spatial filtering to achieve comparably good performance (AUC>0.8 with MLP$_{\emptyset}$). Indeed, these classifiers can provide better results than when using CSP as a preprocessing step. In addition, spatial filtering allows reducing the number of features hence increasing the processing speed for both training and testing the classifier. Although the number of spatial filters and the number of neurons in the hidden layer were selected based on previous studies [26] and were held constant across the methods, the choice of the parameters and the number of available samples for training might influence the overall results.

Our results were observed in the context of difficult tasks, the difficulty of which was driven by the nature of the stimuli. This difficulty was highlighted by the results of Experiment 1, in which the mean AUC of the behavioral performance was 0.886 and classifier performance also differed between the ground truth based on the stimuli and the ground truth based on the behavioral performance. In difficult tasks, the behavioral and neural responses to stimuli from one class may not always be consistent and distinct from the responses to stimuli from the other class. For example, traditional ERP studies have shown that targets that are presented in difficult RSVP tasks may only elicit a P3 when they are consciously reported [58]. In the present context, this suggests that in Experiment 1, where the task was difficult, the missed targets may have elicited a smaller P3, if any at all. Thus, when classification was based on the ground truth, the target class likely included both robust P3s and attenuated P3s (e.g., amplitude and latency variation). The resulting increased within-class variability and reduced interclass variability was a likely source of the lower classifier performance. On the other hand, it is possible that the difference in performance was due to the inclusion of only those trials with the motor response. However, because the signal was bandpassed between 1 and 10.66 Hz, which excluded a large part of the motor related mu rhythm(8–13 Hz), it is unlikely. Although a difference was observed between motor and nonmotor tasks from 550 through 700 ms in [15], the signal was not bandpassed the same way as in this paper.

An additional aspect of the tasks used here is that the target stimuli were very different from trial to trial. For instance, in Experiments 2 and 3, the targets were people to detect was presented at different locations, orientations, and lighting conditions. This variability within the target class can affect both the reaction time for behavioral performance and also the latency of ERP components like the P300 [59]. More importantly, this variability within a class can affect classifier performance [5]. While the tasks used here are different than the classic oddball task that repeats the same target stimulus [6], the RSVP tasks used here may be more relevant for more real-world target detection tasks [18]. For example, in a real RSVP applications for threat detection, the meaning of the target can change over time because of changes in the local target probability within the stimulus sequence. This variability can modulate the ERP characteristics (amplitude and latency) [27]. Therefore, the classifier and the spatial filters should be tuned to model this variability to become invariant to these ERP deformations. CNNs have been already successfully applied in these kinds of situations [45], [49], [60], and the present results generalize this approach to difficult RSVP tasks.

Whereas the obtained AUC with the different methods is consistent with other studies [11], [20], the performance is not optimal. Two main reasons may explain the level of performance. First, the subject's performance is directly related to the visual stimuli, his/her attentional state, and the parameters of the RSVP paradigm (the stimulus onset asynchrony, the interstimulus interval, and the target probability). Different task parameters, (e.g., less noise in the images, slower RSVP rate) may result in better behavioral and classifier performance. Second, the level of performance can be explained by the assumptions made with the methods. For example, it was assumed that the spatial distribution of the ERP stays stable over time. However, this distribution may change as the subject gains practice on the task or experiences fatigue, and incremental methods should be considered to take into account this effect [38]. Although a BCI should be robust against the dynamic fluctuations in brain signals, the choice of a completely adaptive system or an invariant system remains to be determined. For all these reasons, new efficient machine learning techniques and comfortable paradigms that elicit consistent EEG responses over time need to be identified.

Finally, although BCIs have been mainly applied to disabled persons, the results obtained in this paper are relevant for both disabled and healthy users. For example, RSVP paradigms can be used for spelling tasks because they do not require eye movements, which is advantageous for patient groups that have deficits in oculomotor control [61]. In addition, RSVP tasks are also highly relevant for healthy users who, for example, search for target objects in large scenes (e.g., satellite image analysts). However, even though the performance of the classifiers is relatively high, efficient single-trial detection remains a difficult problem. Indeed, further improvements to classification approaches and BCI paradigms that are robust to a variety of testing conditions are required before these systems can be deployed with confidence in real-world scenarios.

VII. Conclusion

In this paper, we have proposed a CNN with training based on the maximization of the AUC for single-trial detection of ERP in three RSVP tasks. We have compared this system with other state-of-the-art methods by decomposing the spatial filtering step from the classification step. The results highlighted the impact of several supervised spatial filtering

methods and their relationships with classifiers. These methods allowed enhancing and reducing the signal features to facilitate the classification of EEG single-trials in a difficult RSVP task. The obtained results suggest that a CNNs can be effective when used for the detection of EEG single-trials as it combines spatial filtering and classification in an united way. In addition, our results show that this strategy can be more efficient than separating the different steps, i.e., spatial filtering and classification. It is an open question for future studies whether other neural network architectures can better detect single-trial ERP responses.

ACKNOWLEDGMENT

The authors would like to thank J. Sato-Reinhold and J. L. Sy for running subjects in Experiment 1.

REFERENCES

[1] S. J. Luck, *A Introduction to the Event-Related Potential Technique*. Cambridge, MA, USA: MIT Press, 2005.

[2] S. J. Luck, G. F. Woodman, and E. K. Vogel, "Event-related potential studies of attention," *Trends Cognit. Sci.*, vol. 4, no. 11, pp. 432–440, 2000.

[3] J. Polich, "Updating P300: An integrative theory of P3a and P3b," *Clin. Neurophysiol.*, vol. 118, no. 10, pp. 2128–2148, 2007.

[4] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[5] K.-R. Müller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz, "Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring," *J. Neurosci. Methods*, vol. 167, no. 1, pp. 82–90, 2008.

[6] L. Farwell and E. Donchin, "Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalograph. Clin. Neurophysiol.*, vol. 70, no. 6, pp. 510–523, 1988.

[7] B. Hong, F. Guo, T. Liu, X. Gao, and S. Gao, "N200-speller using motion-onset visual response," *Clin. Neurophysiol.*, vol. 120, no. 9, pp. 1658–1666, 2009.

[8] A. Rakotomamonjy and V. Guigue, "BCI competition III : Dataset II—ensemble of SVMs for BCI P300 speller," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 3, pp. 1147–1154, Mar. 2008.

[9] H. Cecotti, "Spelling with non-invasive brain-computer interfaces—Current and future trends," *J. Physiol. Paris*, vol. 105, nos. 1–3, pp. 106–114, 2011.

[10] P. Sajda, A. Gerson, and L. Parra, "High-throughput image search via single-trial event detection in a rapid serial visual presentation task," in *Proc. 1st Int. IEEE EMBS Conf. Neural Eng.*, Capri Island, Italy, Mar. 2003, pp. 7–10.

[11] Y. Huang, D. Erdogmus, M. Pavel, S. Mathan, and K. E. Hild, "A framework for visual image search using single-trial brain responses," *Neurocomputing*, vol. 74, pp. 2041–2051, Jun. 2011.

[12] M. C. Potter, "Short-term conceptual memory for pictures," *J. Experim. Psychol., Human Learn. Memory*, vol. 2, no. 5, pp. 509–522, 1976.

[13] M. M. Chun and C. M. Potter, "A two-stage model for multiple target detection in rapid serial visual presentation," *J. Exp. Phys. Human Perception Perform.*, vol. 21, no. 1, pp. 109–127, 1995.

[14] N. Bigdely-Shamlo, A. Vankov, R. R. Ramirez, and S. Makeig, "Brain activity-based image classification from rapid serial visual presentation," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 16, no. 5, pp. 432–441, Oct. 2008.

[15] A. Gerson, L. Parra, and P. Sajda, "Cortically-coupled computer vision for rapid image search," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 174–179, Jun. 2006.

[16] L. C. Parra, C. Christoforou, A. D. Gerson, M. Dyrholm, A. Luo, M. Wagner, *et al.*, "Spatio-temporal linear decoding of brain State: Application to performance augmentation in high-throughput tasks," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 95–115, Jan. 2008.

[17] E. A. Pohlmeyer, D. C. Jangraw, J. Wang, S. F. Chang, and P. Sajda, "Combining computer and human vision into a BCI: Can the whole be greater than the sum of its parts?" in *Proc. 32nd Int. IEEE EMBC Conf.*, Buenos Aires, Argentina, Sep. 2010, pp. 138–141.

[18] E. A. Pohlmeyer, J. Wang, D. C. Jangraw, B. Lou, S. Chang, and P. Sajda, "Closing the loop in cortically-coupled computer vision: A brain-computer interface for searching image databases," *J. Neural Eng.*, vol. 8, no. 3, p. 036025, 2011.

[19] J. Touryan, L. Gibson, H. J. Horne, and P. Weber, "Real-time measurement of face recognition in rapid serial visual representation," *Frontiers Psychol.*, vol. 2, no. 42, pp. 1–8, 2011.

[20] K. E. Hild, M. Kurimo, and V. D. Calhoun, "The sixth annual MLSP competition, 2010," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Kittila, Finland, Sep. 2010, pp. 107–111.

[21] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: Application to brain-computer interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, Aug. 2009.

[22] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, Jan. 2008.

[23] L. Parra, C. Alvino, A. Tang, B. Pearlmutter, N. Yeung, A. Osman, *et al.*, "Single trial detection in EEG and MEG: Keeping it linear," *Neurocomputing*, vols. 52–54, pp. 177–183, Mar. 2003.

[24] U. Hoffmann, J. Vesin, K. Diserens, and T. Ebrahimi, "An efficient P300-based brain-computer interface for disabled subjects," *J. Neurosci. Methods*, vol. 167, no. 1, pp. 115–125, 2008.

[25] B. Labbé, X. Tian, and A. Rakotomamonjy, "MLSP competition, 2010: Description of the third place method," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Kittila, Finland, Sep. 2010, pp. 116–117.

[26] H. Cecotti and A. Gräser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, Mar. 2011.

[27] R. J. Johnson, "A triarchic model of P300 amplitude," *Psychophysiology*, vol. 23, no. 4, pp. 367–384, 1986.

[28] C. J. Gonsalvez and J. Polich, "P300 amplitude is determined by target-to-target interval," *Psychophysiology*, vol. 39, no. 3, pp. 388–396, 2002.

[29] B. Blankertz, M. Kawanabe, R. Tomioka, F. Hohlefeld, V. Nikulin, and K.-R. Müller, "Invariant common spatial patterns: Alleviating nonstationarities in brain-computer interfacing," in *Advances in Neural Information Processing Systems*, vol. 20, Cambridge, MA, USA: MIT Press, 2008, pp. 1–8.

[30] C. Brunner, M. Naeem, R. Leeb, B. Graimann, and G. Pfurtscheller, "Spatial filtering and selection of optimized components in four class motor imagery EEG data using independent components analysis," *Pattern Recognit. Lett.*, vol. 28, no. 8, pp. 957–964, 2007.

[31] G. Pfurtscheller, C. Guger, and H. Ramoser, "EEG-based brain-computer interface using subject-specific spatial filters," in *Proc. Int. Work Conf. Artif. Natural Neural Netw.*, vol. 2. Alicante, Spain, 1999, pp. 248–254.

[32] R. Tomioka, N. J. Hill, B. Blankertz, and K. Aihara, "Adapting spatial filter methods for nonstationary BCIs," in *Proc. Workshop Inf. Based Induction Sci.*, Osaka, Japan, 2006, pp. 1–6.

[33] H. Cecotti, B. Rivet, M. Congedo, C. Jutten, O. Bertrand, E. Maby, *et al.*, "A robust sensor selection method for P300 brain-computer interfaces," *J. Neural Eng.*, vol. 8, no. 1, p. 016001, 2011.

[34] B. Rivet, H. Cecotti, E. Maby, and J. Mattout, "Impact of spatial filters during sensor selection in a visual P300 brain-computer interface," *Brain Topograph.*, vol. 12, no. 1, pp. 55–63, 2012.

[35] B. Rivet and A. Souloumiac, "Optimal linear spatial filters for event-related potentials based on a spatio-temporal model: Asymptotical performance analysis," *Signal Process.*, vol. 93, no. 2, pp. 387–398, 2013.

[36] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD, USA: Johns Hopkins Univ., 1996.

[37] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, "Towards zero training for brain-computer interfacing," *PLoS ONE*, vol. 3, no. 8, p. e2967, 2008.

[38] Q. Zhao, L. Zhang, A. Cichocki, and J. Li, "Incremental common spatial pattern algorithm for BCI," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Hong Kong, Jun. 2008, pp. 2656–2659.

[39] M. Arvaneh, G. Cuntai, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 610–619, Apr. 2013.

[40] H. Zhang, H. Yang, and C. Guan, "Bayesian learning for spatial filtering in an EEG-based brain-computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 7, pp. 1049–1060, Jul. 2013.

[41] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, 1992.

[42] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[43] C.-C. Chang and C.-J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[44] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[45] P. Y. Simard, D. Steinkraus, and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. 7th Int. Conf. Document Anal. Recognit.*, Edinburgh, Scotland, U.K., 2003, pp. 958–962.

[46] R. Hadsell, P. Sermanet, M. Scoffier, A. Erkan, K. Kavackuoglu, U. Muller, *et al.*, "Learning long-range vision for autonomous off-road driving," *J. Field Robot.*, vol. 26, no. 2, pp. 120–144, 2009.

[47] H. Cecotti, "A time-frequency convolutional neural network for the offline classification of steady-state visual evoked potential responses," *Pattern Recognit. Lett.*, vol. 32, no. 8, pp. 1145–1153, 2011.

[48] P. Mirowski, D. Madhavan, Y. LeCun, and R. Kuzniecky, "Classification of patterns of EEG synchronization for seizure prediction," *Clin. Neurophysiol.*, vol. 120, no. 11, pp. 1927–1940, 2009.

[49] Y. Bengio and Y. LeCun, "Scaling learning algorithms towards AI," in *Large-Scale Kernel Machines*, L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, Eds. Cambridge, MA, USA: MIT Press, 2007.

[50] Y. LeCun, L. Bottou, G. Orr, and K. R. Müller, "Efficient backprop," in *Neural Networks: Tricks of the Trade*, G. Orr, Ed. New York, NY, USA: Springer-Verlag, 1998.

[51] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[52] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[53] N. Troje and H. H. Bülthoff, "Face recognition under varying poses: The role of texture and shape," *Vis. Res.*, vol. 36, no. 12, pp. 1761–1771, 1996.

[54] F. Sharbrough, G. Chatrian, and R. P. E. A. Lesser, *Guidelines for Standard Electrode Position Nomenclature*. Bloomfield, IL, USA: Amer. EEG Soc., 1990.

[55] J. M. Leiva and S. M. M. Martens, "MLSP competition, 2010: Description of the first place method," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, Kittila, Finland, Sep. 2010, pp. 112–113.

[56] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in *Advances in Neural Information Processing Systems*. Vancouver, BC, Canada: MIT Press, 2003, pp. 1–8.

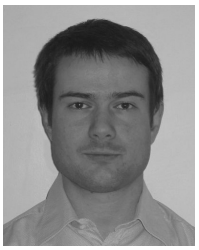
[57] R. Tomioka and K. R. Müller, "A regularized discriminative framework for EEG analysis with application to brain-computer interface," *Neuroimage*, vol. 49, no. 1, pp. 415–432, 2010.

[58] E. K. Vogel, S. J. Luck, and K. L. Shapiro, "Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink," *J. Experim. Psychol., Human Perception Perform.*, vol. 24, no. 6, pp. 1656–1674, 1998.

[59] M. Kutas, G. McCarthy, and E. Donchin, "Augmenting mental chronometry: The p300 as a measure of stimulus evaluation time," *Science*, vol. 197, no. 4305, pp. 792–795, 1977.

[60] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, 2007, pp. 473–480.

[61] L. Acqualagna, M. S. Treder, M. Schreuder, and B. Blankertz, "A novel brain-computer interface based on the rapid serial visual presentation paradigm," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* , Buenos Aires, Argentina, Sep. 2010, pp. 2686–2689.



Hubert Cecotti received the M.Sc. and Ph.D. degrees in computer science from the University of Lorraine, Lorraine, France, in 2002 and 2005, respectively.

He was a Lecturer in computer science with the University Henri Poincare and ESIAL, Nancy, France, in 2006 and 2007. From 2008 to 2013, he was successively a Research Scientist with the Institute of Automation, Bremen University, Bremen, Germany, the Gipsa-Lab CNRS, Grenoble, France, and the University of California Santa Barbara, Santa Barbara, CA, USA, where he was involved in electroencephalography signal processing and machine learning. He is currently a Lecturer with the School of Computing and Intelligent Systems, University of Ulster, Londonderry, U.K. His current research interests include neural networks, pattern recognition, and brain–computer interfaces.

Miguel P. Eckstein received the B.S. degree in physics and psychology from UC Berkeley, and the Ph.D. degree in cognitive psychology from University of California, Los Angeles, CA.

He was with the Department of Medical Physics and Imaging, Cedars Sinai Medical Center and NASA Ames Research Center before moving to UC Santa Barbara. He has published over 120 articles relating to computational human vision, visual attention and search, perceptual learning, and the perception of medical images.

Dr. Eckstein has served as the Chair of the Vision Technical Group of the Optical Society of America, the Chair of the Human Performance, Image Perception and Technology Assessment Conference of the Society for Optical Engineering (SPIE) Medical Imaging Annual Meeting, and as a member of various National Institute of Health study section panels. He served as the Vision Editor of the *Journal of the Optical Society of America A* from 2005 to 2001, and is currently on the board of editors of the *Journal of Vision*, and the board of directors of the Vision Sciences Society. He is a recipient of the Optical Society of America Young Investigator Award, the Society for Optical Engineering Image Perception Cum Laude Award, the Cedars Sinai Young Investigator Award, the National Science Foundation CAREER Award, and the National Academy of Sciences Troland Award.

Barry Giesbrecht received the B.A. degree in psychology from the University of Waterloo, Waterloo, Canada, and the Ph.D. degree in psychology from the University of Alberta, Edmonton, AB, Canada, in 1995 and 1999, respectively.

He did the post-doctoral training from the Center for Cognitive Neuroscience, Duke University, Durham, NC, USA, from 1999 to 2002, and from the Center for Mind and Brain, University of California, Davis, CA, USA, from 2002 to 2004. In 2004, he joined the Department of Psychological and Brain Sciences, University of California, Santa Barbara, CA, where he is currently an Associate Professor. His current research interests include understanding the cognitive and neural bases of the human attention system, in particular, identifying those mechanisms that control attention in a variety contexts and the fundamental limitations of these mechanisms. To investigate these issues, he uses a combination of methods, including measuring behavioral performance in cognitive psychological tasks, measurements of brain activity acquired using electroencephalography and functional magnetic resonance imaging.